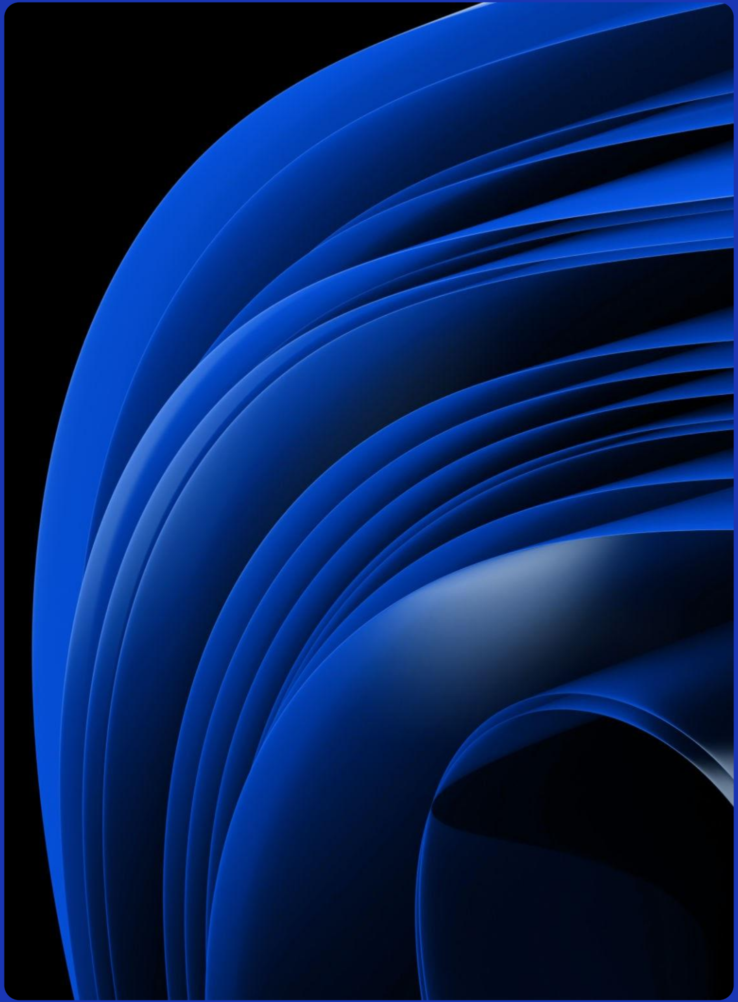
Miri Eisenberg

DS 210 - Introduction to Data Science

Final Project

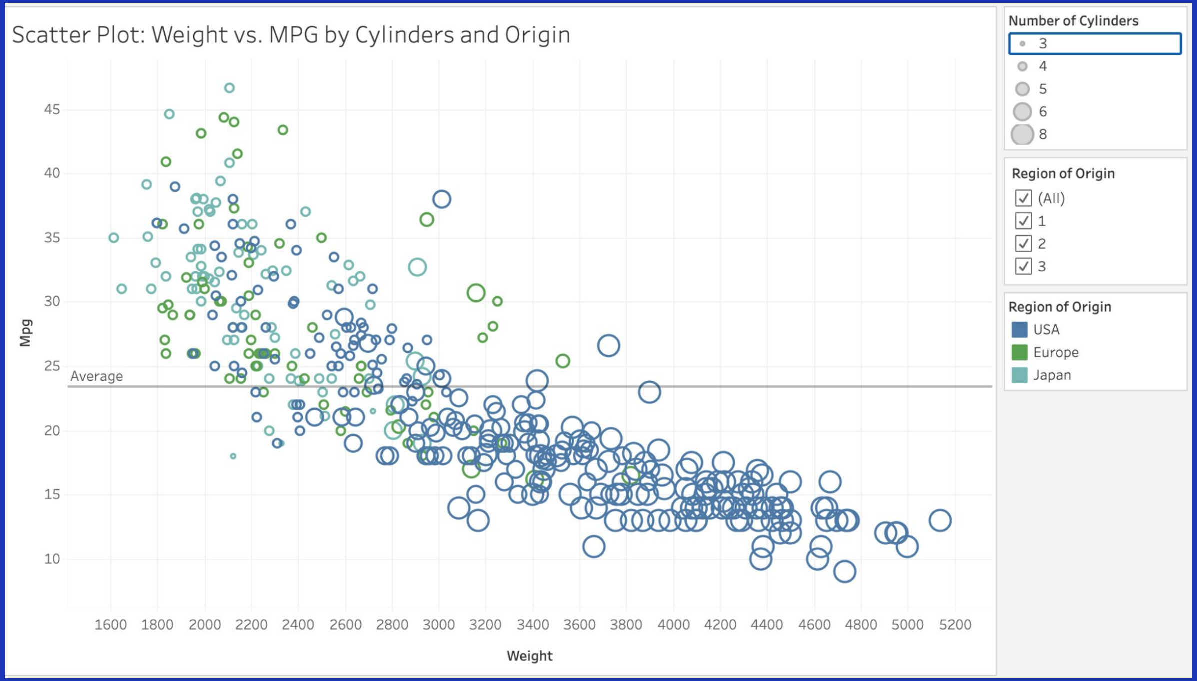
Confidential

Copyright ©



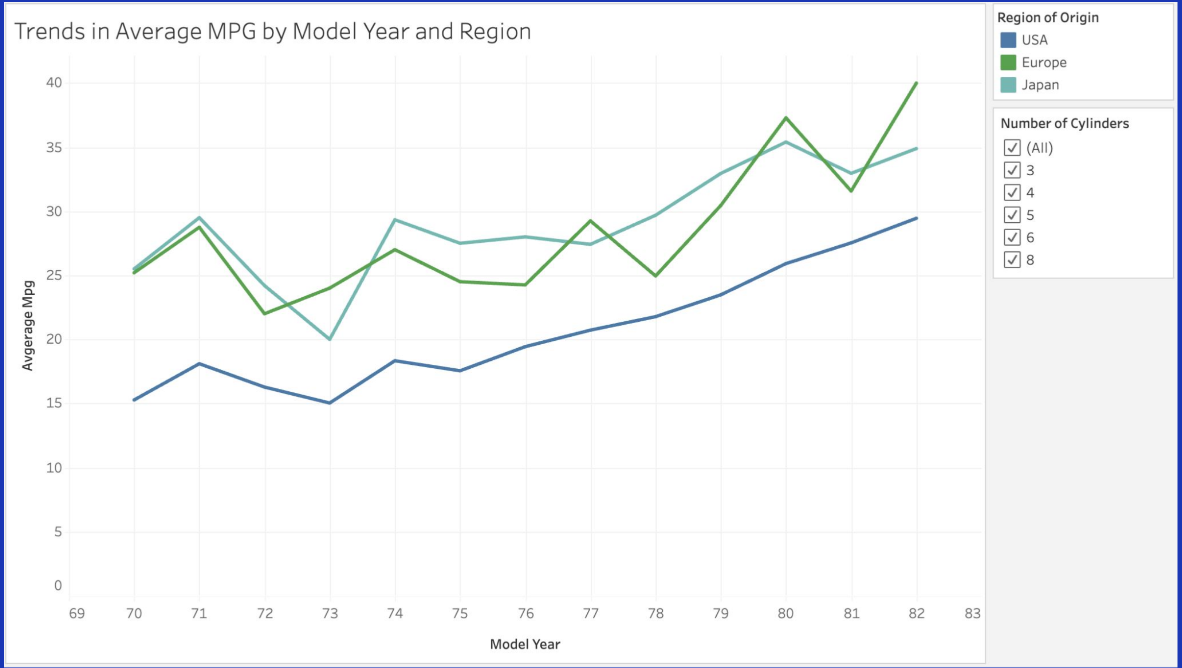
Part I

This visualization shows the relationship between car weight and fuel efficiency MPG. It also shows the differences by region of origin (USA, Europe, Japan) and engine size (number of cylinders). Lighter cars generally have higher MPG, for example Japanese and European cars are more fuel-efficient and lighter. American cars are usually heavier with lower MPG. The size of the circles represents the number of cylinders, showing that cars with larger engines - more cylinders are typically heavier and less efficient. The dropdown filters allow you to sort by origin and size of cylinders.



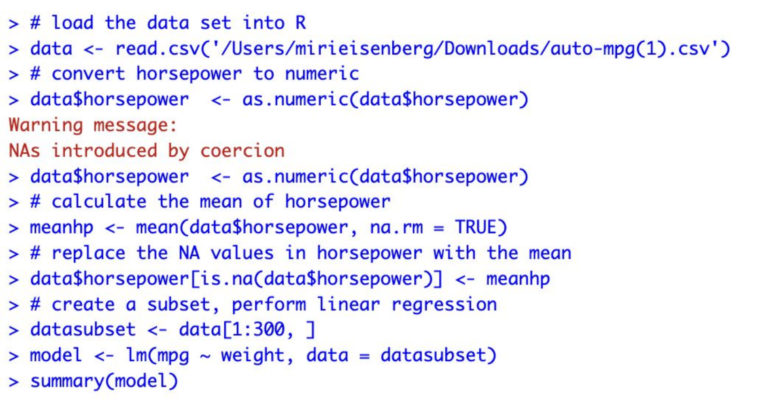
https://public.tableau.com/views/DS210/Sheet1?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

This visualization shows how the average MPG has changed over time (by model year) for cars from different regions - USA, Europe, and Japan. It uses line charts to compare trends, revealing regional differences in fuel efficiency improvements. You can see if Japanese cars improved faster than American cars or if certain years shifted. The visualization shows how advancements in car design and regulations impacted fuel efficiency across regions. The filter allows you to see where the engine size changed over time. This visual shows how MPG has improved over the last few years.



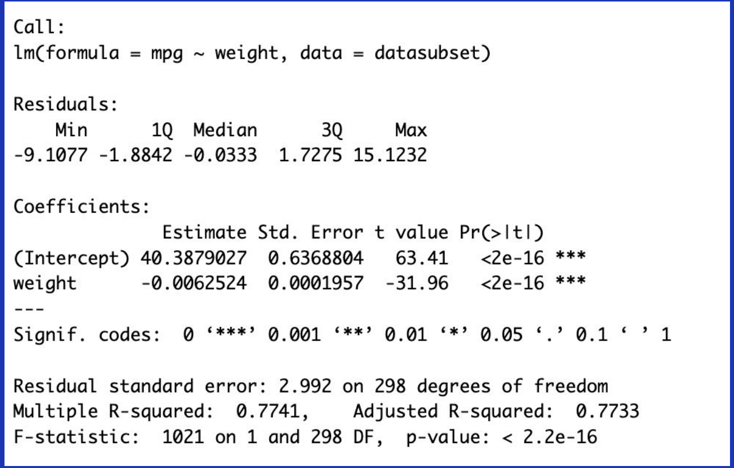
https://public.tableau.com/views/VIS1_17348386851000/Sheet2?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

CODE

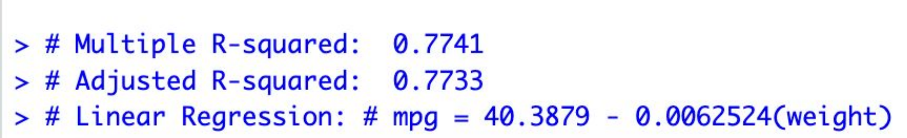


```
> # load the data set into R
> data <- read.csv('/Users/mirieisenberg/Downloads/auto-mpg(1).csv')
> # convert horsepower to numeric
> data$horsepower  <- as.numeric(data$horsepower)
Warning message:
NAs introduced by coercion
> data$horsepower  <- as.numeric(data$horsepower)
> # calculate the mean of horsepower
> meanhp <- mean(data$horsepower, na.rm = TRUE)
> # replace the NA values in horsepower with the mean
> data$horsepower[is.na(data$horsepower)] <- meanhp
> # create a subset, perform linear regression
> datasubset <- data[1:300, ]
> model <- lm(mpg ~ weight, data = datasubset)
> summary(model)
```

RESULT

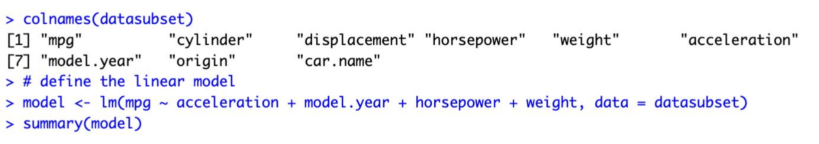


```
Call:
lm(formula = mpg ~ weight, data = datasubset)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1077 -1.8842 -0.0333  1.7275 15.1232 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.3879027  0.6368804   63.41   <2e-16 ***
weight      -0.0062524  0.0001957  -31.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.992 on 298 degrees of freedom
Multiple R-squared:  0.7741,	Adjusted R-squared:  0.7733 
F-statistic:  1021 on 1 and 298 DF,  p-value: < 2.2e-16
```



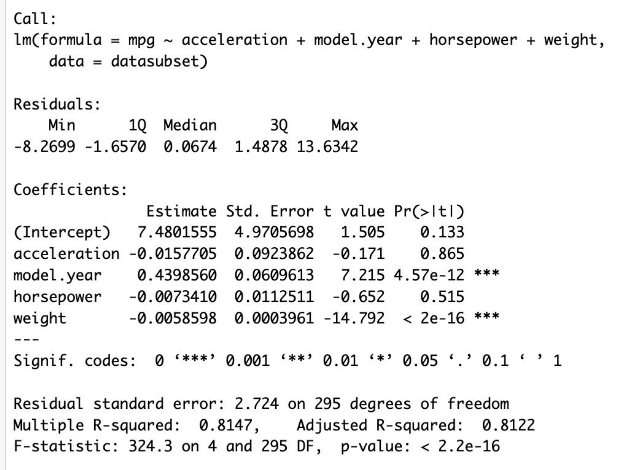
```
> # Multiple R-squared:  0.7741
> # Adjusted R-squared:  0.7733
> # Linear Regression: # mpg = 40.3879 - 0.0062524(weight)
```

CODE

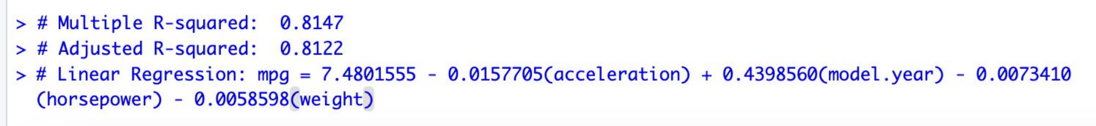


```
> colnames(datasubset)
[1] "mpg"          "cylinder"     "displacement" "horsepower"   "weight"       "acceleration"
[7] "model.year"   "origin"       "car.name"
> # define the linear model
> model <- lm(mpg ~ acceleration + model.year + horsepower + weight, data = datasubset)
> summary(model)
```

RESULT

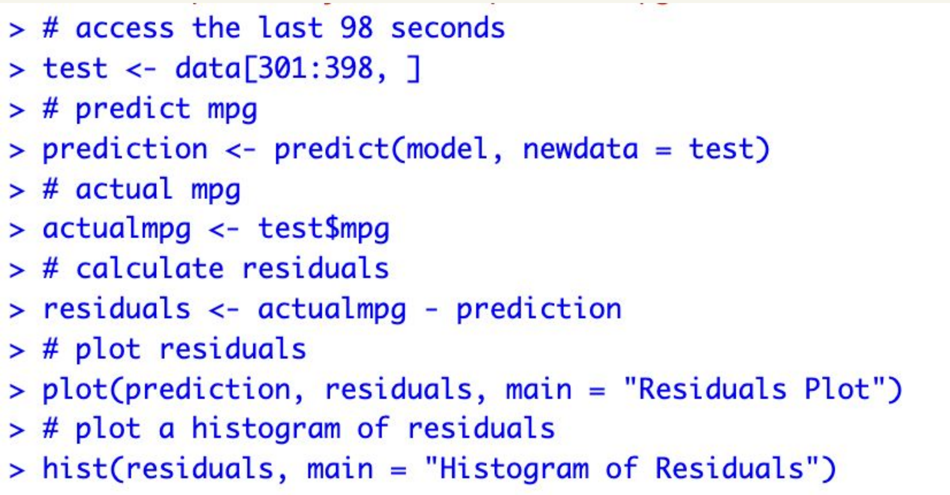


```
Call:
lm(formula = mpg ~ acceleration + model.year + horsepower + weight,
    data = datasubset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2699 -1.6570  0.0674  1.4878 13.6342

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4801555  4.9705698   1.505    0.133
acceleration -0.0157705  0.0923862  -0.171    0.865
model.year    0.4398560  0.0609613   7.215 4.57e-12 ***
horsepower   -0.0073410  0.0112511  -0.652    0.515
weight       -0.0058598  0.0003961 -14.792  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.724 on 295 degrees of freedom
Multiple R-squared:  0.8147,	Adjusted R-squared:  0.8122
F-statistic: 324.3 on 4 and 295 DF,  p-value: < 2.2e-16
```

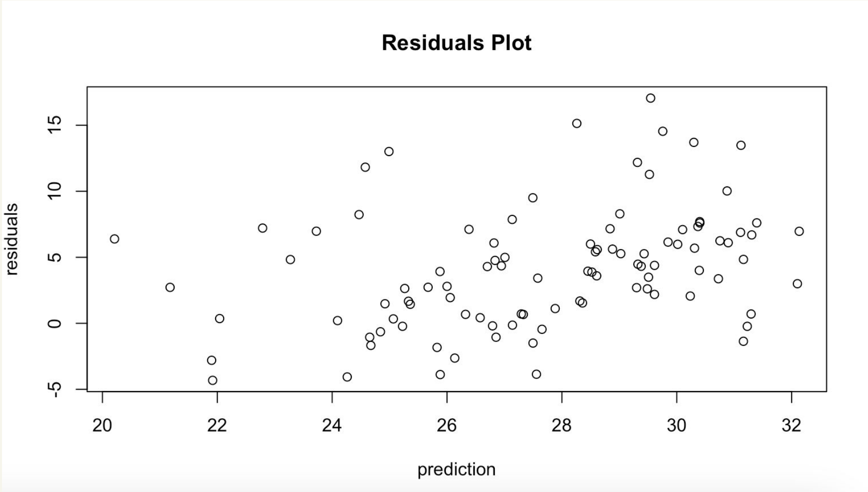


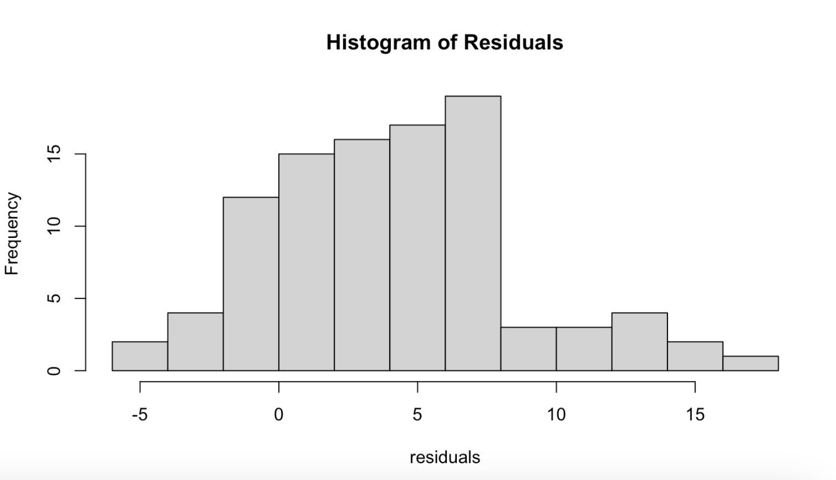
```
> # Multiple R-squared:  0.8147
> # Adjusted R-squared:  0.8122
> # Linear Regression: mpg = 7.4801555 - 0.0157705(acceleration) + 0.4398560(model.year) - 0.0073410
  (horsepower) - 0.0058598(weight)
```

Residuals

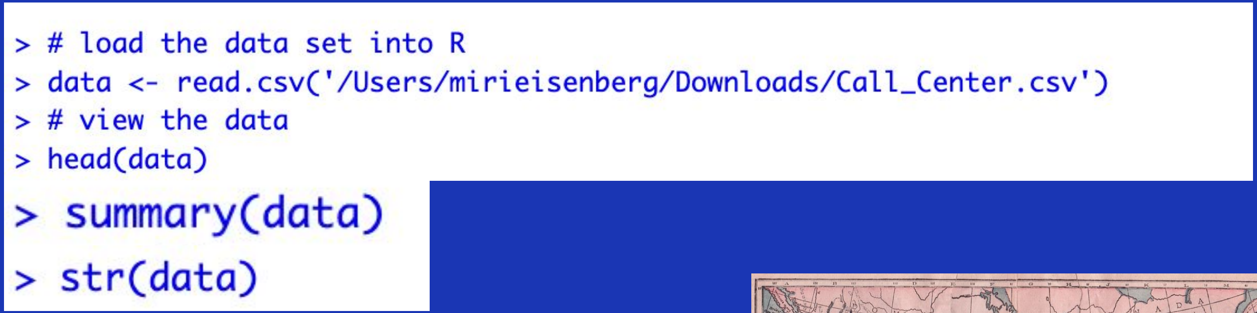


```
> # access the last 98 seconds
> test <- data[301:398, ]
> # predict mpg
> prediction <- predict(model, newdata = test)
> # actual mpg
> actualmpg <- test$mpg
> # calculate residuals
> residuals <- actualmpg - prediction
> # plot residuals
> plot(prediction, residuals, main = "Residuals Plot")
> # plot a histogram of residuals
> hist(residuals, main = "Histogram of Residuals")
```



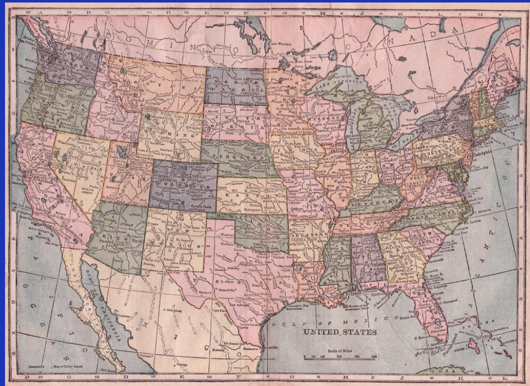


Part II



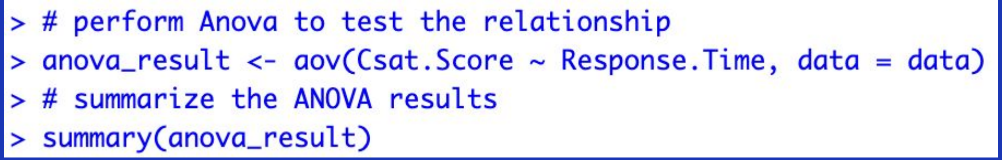
```
> # load the data set into R
> data <- read.csv('/Users/mirieisenberg/Downloads/Call_Center.csv')
> # view the data
> head(data)
> summary(data)
> str(data)
```

Loading The Data



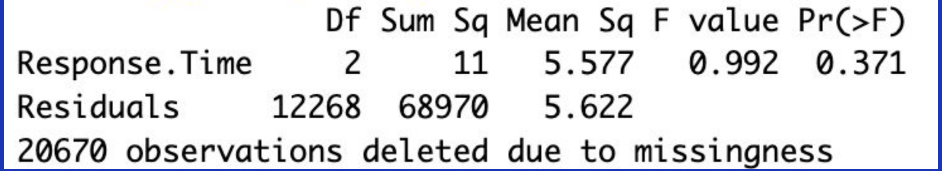
What is the relationship between response time and customer satisfaction (CSAT)?

CODE



```
> # perform Anova to test the relationship
> anova_result <- aov(Csat.Score ~ Response.Time, data = data)
> # summarize the ANOVA results
> summary(anova_result)
```

RESULT

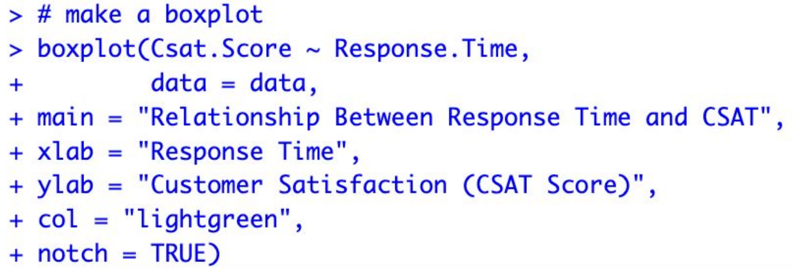


```
                 Df Sum Sq Mean Sq F value Pr(>F)
Response.Time     2     11   5.577   0.992  0.371
Residuals     12268  68970   5.622
20670 observations deleted due to missingness
```

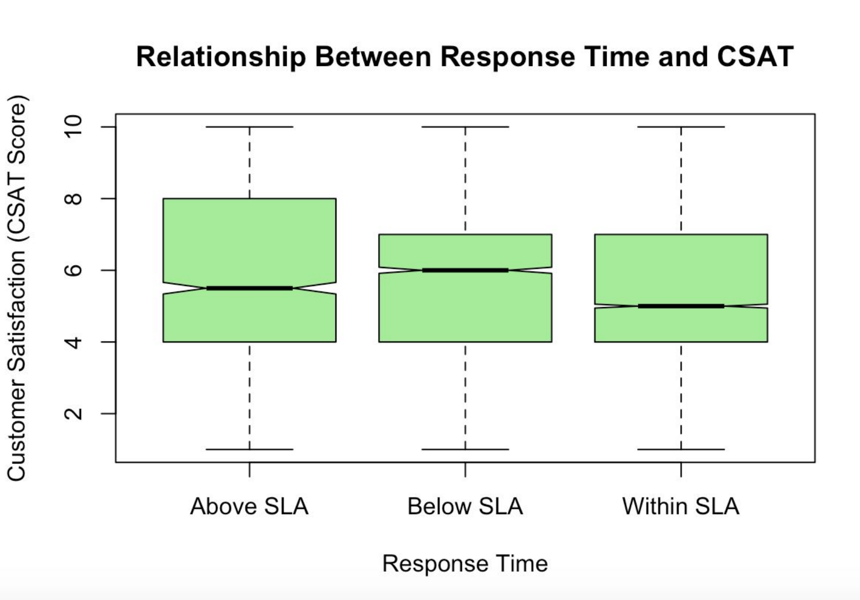
P-value = 0.371. This number is greater than 0.05 which means that we fail to reject the null hypothesis. Practically speaking this means that there is no strong evidence between the Response time and customer satisfaction. If the p-value would have been less than 0.05 we would see a strong correlation between response time and customer satisfaction.

Df (degrees of freedom) = 2. This means that there are 3 levels/categories to response time - we find this by using the equation n-1.

VISUALIZATION



```
> # make a boxplot
> boxplot(Csat.Score ~ Response.Time,
+         data = data,
+ main = "Relationship Between Response Time and CSAT",
+ xlab = "Response Time",
+ ylab = "Customer Satisfaction (CSAT Score)",
+ col = "lightgreen",
+ notch = TRUE)
```



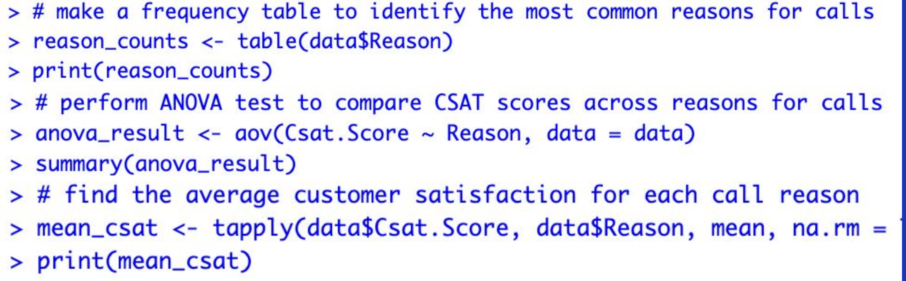
The Medians are similar in all three groups, showing that all groups have similar customer satisfaction scores.

This boxplot does not have any outliers showing consistent responses.

The boxes are similar sizes (above SLA is a little larger) showing that the range of Customer Satisfaction scores are similar throughout the group. This further shows that response time does not affect customer satisfaction.

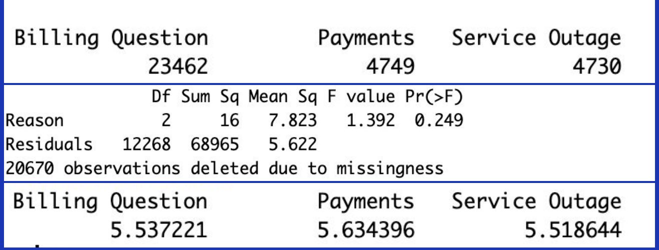
What are the most common reasons for calls, and how do they impact CSAT?

CODE



```
> # make a frequency table to identify the most common reasons for calls
> reason_counts <- table(data$Reason)
> print(reason_counts)
> # perform ANOVA test to compare CSAT scores across reasons for calls
> anova_result <- aov(Csat.Score ~ Reason, data = data)
> summary(anova_result)
> # find the average customer satisfaction for each call reason
> mean_csat <- tapply(data$Csat.Score, data$Reason, mean, na.rm =
> print(mean_csat)
```

RESULT



```
Billing Question          Payments   Service Outage
           23462              4749             4730
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
Reason          2     16   7.823   1.392  0.249
Residuals   12268  68965   5.622
20670 observations deleted due to missingness
```

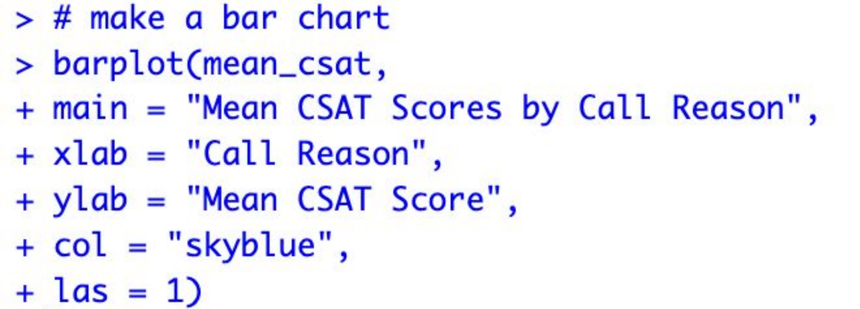
```
Billing Question          Payments   Service Outage
        5.537221          5.634396         5.518644
```

The most common calls relate to Billing Questions, and the least common calls relate to service outage

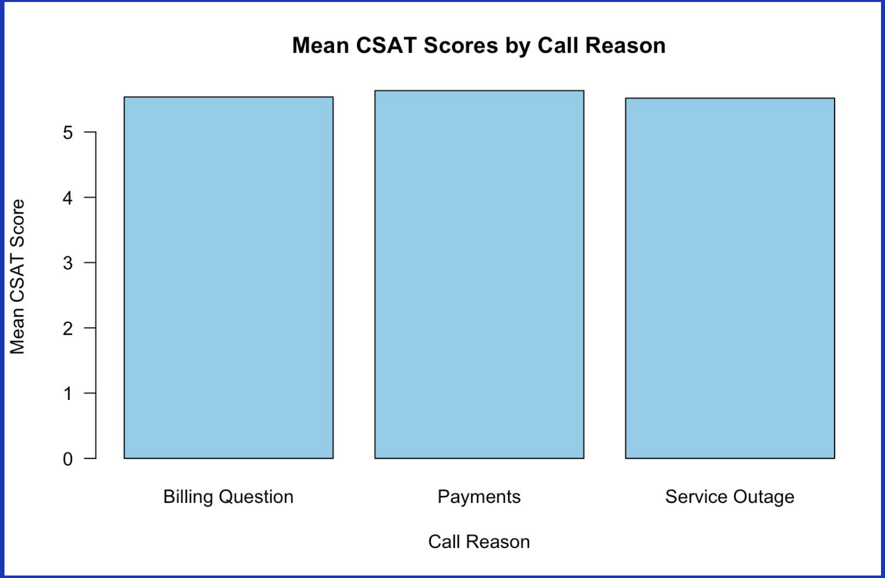
P-value = 0.249. This number is greater than 0.05 which means that we fail to reject the null hypothesis. Practically speaking this means that there is no strong evidence between the Customer satisfaction and reasons for calls. If the p-value would have been less than 0.05 we would see a strong correlation between customer satisfaction and reasons for calls.

The means are similar to each other, strengthening the conclusion that reasons for calls do not impact customer satisfaction.

VISUALIZATION



```
> # make a bar chart
> barplot(mean_csat,
+ main = "Mean CSAT Scores by Call Reason",
+ xlab = "Call Reason",
+ ylab = "Mean CSAT Score",
+ col = "skyblue",
+ las = 1)
```



The heights of the bars are almost exactly the same three groups, showing that there is only a very slight difference in customer satisfaction scores between the reasons for calls.

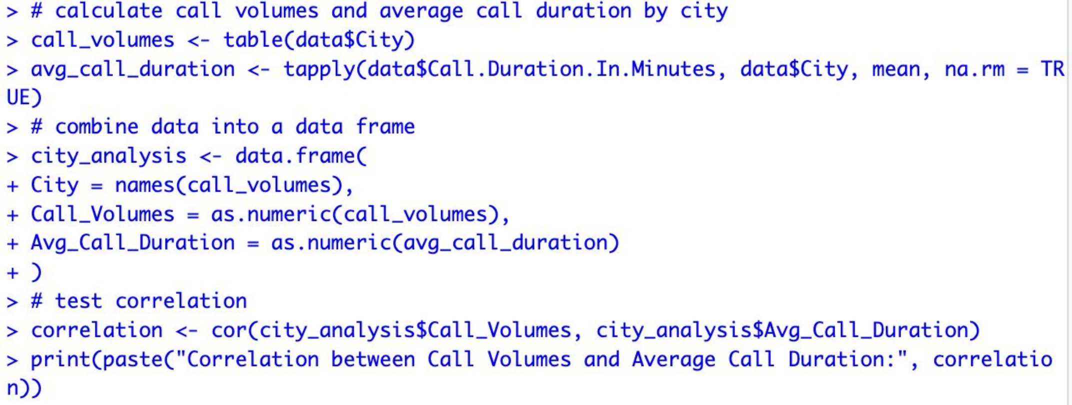
There is no meaningful correlation.

The overall customer satisfaction score is around 5.5 which shows that customers are neutral or slightly unsatisfied.

This bar graph does not have any outliers.

Is there a correlation between the number of calls received in a city and the average call duration for that city?

CODE

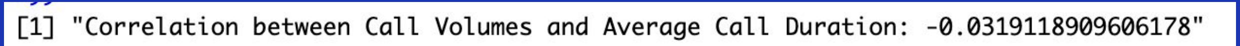


```
> # calculate call volumes and average call duration by city
> call_volumes <- table(data$City)
> avg_call_duration <- tapply(data$Call.Duration.In.Minutes, data$City, mean, na.rm = TR
UE)
> # combine data into a data frame
> city_analysis <- data.frame(
+ City = names(call_volumes),
+ Call_Volumes = as.numeric(call_volumes),
+ Avg_Call_Duration = as.numeric(avg_call_duration)
+ )
> # test correlation
> correlation <- cor(city_analysis$Call_Volumes, city_analysis$Avg_Call_Duration)
> print(paste("Correlation between Call Volumes and Average Call Duration:", correlatio
n))
```

The correlation coefficient is -0.0319118909606178 which is close to 0. This indicates that there is no meaningful correlation between call volumes and average call duration. This proves that the number of calls a city handles does not significantly affect how long the calls last.

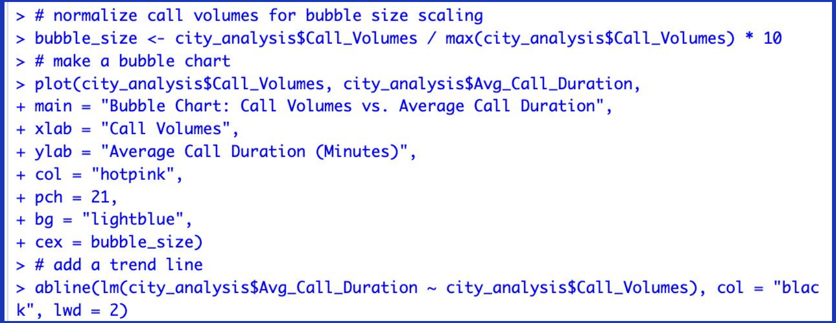
RESULT

The slight negative value means that as call volumes increase, average call durations may slightly decrease, but this relationship is insignificant.

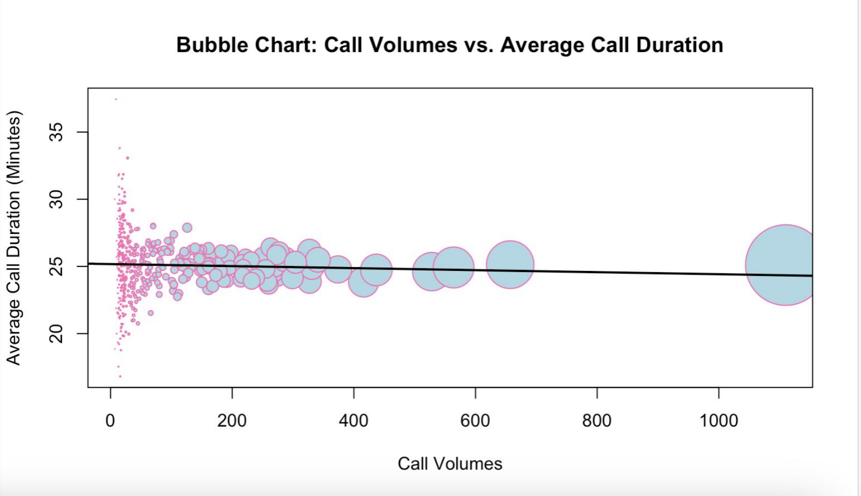


```
[1] "Correlation between Call Volumes and Average Call Duration: -0.0319118909606178"
```

VISUALIZATION



```
> # normalize call volumes for bubble size scaling
> bubble_size <- city_analysis$Call_Volumes / max(city_analysis$Call_Volumes) * 10
> # make a bubble chart
> plot(city_analysis$Call_Volumes, city_analysis$Avg_Call_Duration,
+ main = "Bubble Chart: Call Volumes vs. Average Call Duration",
+ xlab = "Call Volumes",
+ ylab = "Average Call Duration (Minutes)",
+ col = "hotpink",
+ pch = 21,
+ bg = "lightblue",
+ cex = bubble_size)
> # add a trend line
> abline(lm(city_analysis$Avg_Call_Duration ~ city_analysis$Call_Volumes), col = "blac
k", lwd = 2)
```

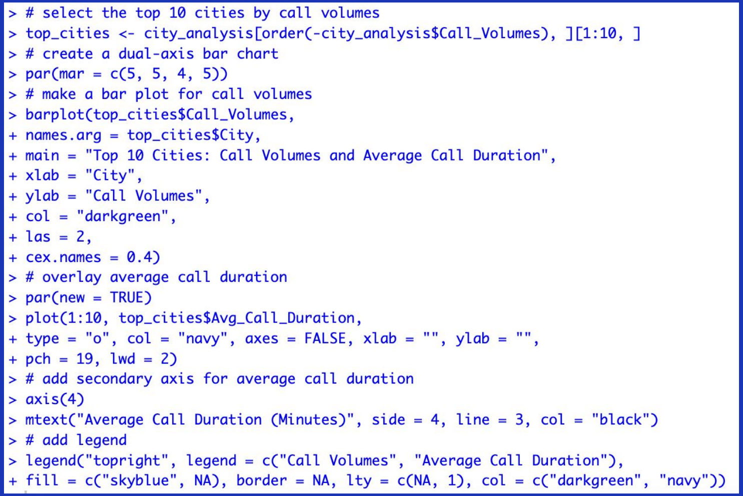


The slight slight downward slope of the trend line suggests a weak correlation between call volumes and average call duration. This aligns with the correlation coefficient, which indicates no significant linear correlation.

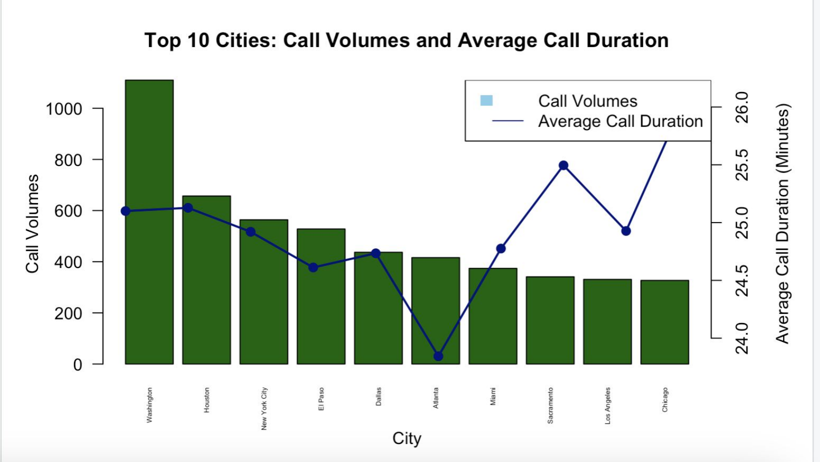
The size of the bubbles is proportional to the call volumes in each city. The bigger the bubble, the more calls a city gets.

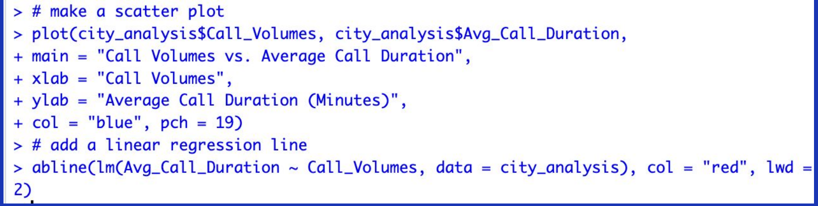
The X-axis represents the total call volumes for each city. The Y-axis represents the average call duration in minutes. The black trend line shows the relationship between the two variables.

The city with around 1,000 calls sticks out however the duration of the calls which are around 25 minutes is the same as the cities with fewer calls.



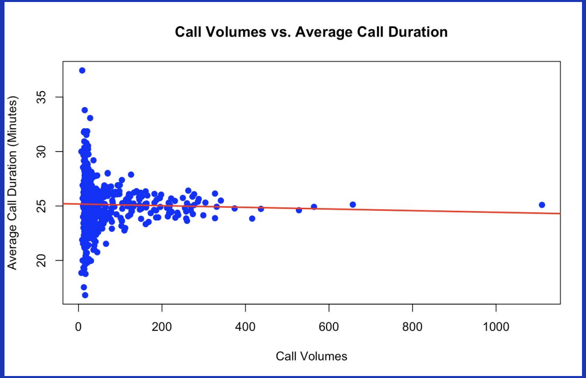
```
> # select the top 10 cities by call volumes
> top_cities <- city_analysis[order(-city_analysis$Call_Volumes), ][1:10, ]
> # create a dual-axis bar chart
> par(mar = c(5, 5, 4, 5))
> # make a bar plot for call volumes
> barplot(top_cities$Call_Volumes,
+ names.arg = top_cities$City,
+ main = "Top 10 Cities: Call Volumes and Average Call Duration",
+ xlab = "City",
+ ylab = "Call Volumes",
+ col = "darkgreen",
+ las = 2,
+ cex.names = 0.4)
> # overlay average call duration
> par(new = TRUE)
> plot(1:10, top_cities$Avg_Call_Duration,
+ type = "o", col = "navy", axes = FALSE, xlab = "", ylab = "",
+ pch = 19, lwd = 2)
> # add secondary axis for average call duration
> axis(4)
> mtext("Average Call Duration (Minutes)", side = 4, line = 3, col = "black")
> # add legend
> legend("topright", legend = c("Call Volumes", "Average Call Duration"),
+ fill = c("skyblue", NA), border = NA, lty = c(NA, 1), col = c("darkgreen", "navy"))
```





```
> # make a scatter plot
> plot(city_analysis$Call_Volumes, city_analysis$Avg_Call_Duration,
+ main = "Call Volumes vs. Average Call Duration",
+ xlab = "Call Volumes",
+ ylab = "Average Call Duration (Minutes)",
+ col = "blue", pch = 19)
> # add a linear regression line
> abline(lm(Avg_Call_Duration ~ Call_Volumes, data = city_analysis), col = "red", lwd =
2)
```

The scatter plot shows no clear trend, with call durations remaining stable regardless of call volumes.



Washington, Houston, and New York City have high call volumes with consistent call durations proving efficiency.
Atlanta has a higher average duration despite low call volumes, suggesting inefficiencies or complex calls.

Optimizing Businesses

What is the relationship between response time and customer satisfaction (CSAT)?

Since response time has little impact on CSAT, focus should shift to resolving issues effectively on the first interaction by improving agent behavior, communication, and resolution clarity. SLA targets can be adjusted to optimize efficiency and reduce costs without affecting satisfaction.

What are the most common reasons for calls, and how do they impact CSAT?

Improve processes across all call types by addressing systemic issues like resolution quality, communication, and agent training. Prioritize proactive communication for Service Outages and provide targeted training to handle Billing Questions and Payments more effectively.

Is there a correlation between the number of calls received in a city and the average call duration for that city?

Use high-volume cities with stable durations as benchmarks for best practices. Address inefficiencies in low-volume cities with higher durations through process improvements or training. Base staffing decisions on call complexity and type, not just volume.

Part III

Github